CRM 86-231 / November 1986

ON SOME ISSUES IN THE ACCELERATED CAT-ASVAB PROJECT

D. R. Divgi

Work conducted under contract N00014-83-C-0725.

This Research Memorandum represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
CRM 86-231

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Center for Naval Analyses

6b. OFFICE SYMBOL *(If applicable)*
CNA

7a. NAME OF MONITORING ORGANIZATION
Commandant of the Marine Corps (Code RDS)

6c. ADDRESS *(City, State, and ZIP Code)*
4401 Ford Avenue
Alexandria, Virginia 22302-0268

7b. ADDRESS *(City, State, and ZIP Code)*
Headquarters, Marine Corps
Washington, D.C. 20380

8a. NAME OF FUNDING / ORGANIZATION
Office of Naval Research

8b. OFFICE SYMBOL *(If applicable)*
ONR

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-83-C-0725

8c. ADDRESS *(City, State, and ZIP Code)*
800 North Quincy Street
Arlington, Virginia 22217

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
65153M	C0031		

11. TITLE *(Include Security Classification)*
On Some Issues in the Accelerated CAT-ASVAB Project

12. PERSONAL AUTHOR(S)
D. R. Divgi

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM TO

14. DATE OF REPORT *(Year, Month, Day)*
November 1986

15. PAGE COUNT
36

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The Accelerated CAT-ASVAB Project (ACAP) may serve as the pilot version for national implementation of computerized adaptive testing (CAT) for the Armed Services Vocational Aptitude Battery (ASVAB). Two major decisions in ACAP involve the introduction of new items into the tests, and setting time limits. This Research Memorandum takes the position that the long-term benefits which CAT may provide are more important than purely technical concerns and makes recommendations based on this position.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Lt.Col. G. W. Russell

22b. TELEPHONE *(Include Area Code)*
(202) 694-3491

22c OFFICE SYMBOL
RDS-40

83 APR edition may be used until exhausted
All other editions are obsolete

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

7 November 1986

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 86-231

Encl: (1) CNA Research Memorandum 86-231, "On Some Issues in the Accelerated CAT-ASVAB Project," by D. R. Divgi, November 1986

1. Enclosure (1) is forwarded as a matter of possible interest.

2. The Department of Defense (DOD) may implement a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) in the near future. The Accelerated CAT-ASVAB Project (ACAP) is the pilot project for possible national implementation of CAT-ASVAB. This Research Memorandum discusses some implications of the need to maximize the benefit of CAT to DOD, and to use ACAP as the field trial of the methodology which may be used in national implementation.

William H. Sims
Director, Manpower and Training Program
Marine Corps Operations
Analysis Group

Distribution List:
Reverse Page

Subj: Center for Naval Analyses Research Memorandum 86-231

Distribution List

SNDL	
A1	ASSTSECNAV MRA
A1	DASN - MANPOWER (2 copies)
A6	HQMC MPR
	Attn: Deputy Chief of Staff for Manpower (2 copies)
	Attn: Director, Personnel Procurement Division (2 copies)
	Attn: Director, Manpower Plans and Policy Division (2 copies)
	Attn: Director, Personnel Management Division (2 copies)
A6	HQMC TRNG (2 copies)
A6	HQMC RD&S (2 copies)
A6	HQMC RA (2 copies)
A6	HQMC AVN (2 copies)
E3D1	CNR
E3D5	NAVPERSRANDCEN
	Attn: Director, Manpower and Personnel Laboratory
	Attn: Technical Library
	Attn: CAT/ASVAB PMO
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FT1	CNET
V12	CG MCDEC

OPNAV

OP-01
OP-11
OP-12
OP-13

Other

Joint Service CAT-ASVAB Working Group (15 copies)

CRM 86-231 / November 1986

ON SOME ISSUES IN THE ACCELERATED CAT-ASVAB PROJECT

D. R. Divgi

Marine Corps Operations Analysis Group

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

The Accelerated CAT-ASVAB Project (ACAP) may serve as the pilot version for national implementation of computerized adaptive testing (CAT) for the Armed Services Vocational Aptitude Battery (ASVAB). Two major decisions in ACAP involve the introduction of new items into the tests, and setting time limits. This Research Memorandum takes the position that the long-term benefits which CAT may provide are more important than purely technical concerns and makes recommendations based on this position.

[illegible] on site

Dwg #1 [illegible] Sept 07 03

EXECUTIVE SUMMARY

INTRODUCTION

Within a few years the Department of Defense (DOD) may begin administering the Armed Services Vocational Aptitude Battery (ASVAB) using computerized adaptive testing (CAT). Current work on CAT-ASVAB is a part of the Accelerated CAT-ASVAB Project (ACAP). ACAP will provide information about the feasibility of eventual full-scale development (FSD) for nationwide implementation of CAT-ASVAB.

On-line calibration of new items requires "seeding," i.e., inserting them unobtrusively into the operational CAT test. The number of seeded items is a major unresolved issue in ACAP.

Another issue concerns the time allowed for CAT-ASVAB subtests. The current time limits are liberal. If they are reduced, some time will be freed for adding tests to measure new skills, but some applicants may be unable to complete the tests.

The third topic discussed here is the importance of new tests. Such tests will play a major role in the economic benefit which introduction of CAT will provide for DOD.

The position taken in this Research Memorandum is that the primary consideration is the role of ACAP as a pilot project for FSD, and that anything important which is undesirable in FSD should not be done during ACAP.

Discussion of time limits is based on analyses of three data sets: (i) Navy recruits tested in a research study by the Navy Personnel Research and Development Center in 1979; (ii) recruits tested under operational conditions during the norming of ASVAB forms 8, 9 and 10; and (iii) applicants who took form 8a during its initial operational test and evaluation.

IMPROVING SELECTION AND CLASSIFICATION

Adaptive tests can achieve better reliabilities than paper-pencil (PP) tests. However, such increase is of little practical use. Selection and classification of recruits are based on composite scores. Composite reliabilities are already so high that there is very little scope for improving validities through more precise measurement. This point is illustrated in table I using the four High School Occupational Composites: Mechanical and Crafts (MC), Business and Clerical (BC), Electrical and Electronics (EE), and Health, Social & Technology (HST). The first two rows of numbers contain the reliabilities and typical validities in PP ASVAB. The next two rows show what

these will become if each PP subtest is made longer by a half. It is clear that validities will not increase by much even if the precision of CAT scores is equivalent to making PP longer by 50 percent.

TABLE I

GAINS IN RELIABILITY AND VALIDITY WITH
50 PERCENT INCREASE IN TEST LENGTH

	Composite			
	MC	BC	EE	HST
Reliability: PP	.93	.93	.94	.95
Validity: PP	.48	.45	.48	.49
Reliability: PP+50%	.953	.953	.960	.967
Validity: PP+50%	.486	.456	.485	.494

Appreciable gains in validities can be achieved only by introducing new tests which measure skills not currently measured by the ASVAB. This can be done if CAT is designed to save time while equaling rather than exceeding PP ASVAB in reliability.

NUMBER OF SEEDED ITEMS

About a million applicants take the ASVAB each year. Even if only one seeded item per subtest is administered to each applicant in FSD, it will provide 250 new calibrated items every year for each subtest. Seeding more items would waste examinee time which should be used to administer new tests.

Since only one seeded item per subtest suffices during FSD, the same should be done during ACAP. The ACAP sample size will allow on-line calibration of four or five items per subtest. This is enough for making sure that the computer programs work properly.

TIME LIMITS

The present time limits for CAT subtests are based on the rate of test completion by Army recruits on the experimental CAT-ASVAB battery, in which examinees were permitted to work at their own pace. These data almost certainly yield inflated estimates. PP ASVAB data on applicants and recruits were analyzed to examine the issue. The results show the following.

(i) At any given ability level, applicants, being motivated to complete the test and score high, work faster than recruits. (ii) Recruits work faster with specified time limits than they do in a research study conducted to determine time limits. (iii) Those who are unable to complete the test tend to be in the lower mental categories. Therefore the items they will receive in a CAT subtest will be easier than those near the end of a PP ASVAB subtest, which will help them complete the test. (iv) Among applicants, mean numbers of unreached items on PP ASVAB subtests are below 0.8 down to mental category IVa.

An Apple III computer was used in the experimental CAT-ASVAB. The Hewlett-Packard computer to be used during ACAP has a better display screen, which makes it easier to read the items. This will reduce the amount of time needed per item.

The choice of time limits is less a technical issue than a value judgment. It depends on what will be done with the time saved by replacing PP ASVAB with CAT.

IMPORTANCE OF NEW PREDICTORS

As shown in table I, substantially more accurate selection and classification are not possible except by adding new predictors. Therefore the value of CAT to DOD depends heavily on the extent to which additional subtests increase validities of the AFQT and of other composites. Values of these incremental validities are fundamental to any economic analysis of the benefits of CAT. Close cooperation among the services will be needed to avoid unnecessary delays in validating the new tests, and in redefining composites to include these tests.

Operational use of new predictors will require norms for these tests. As they will measure traits quite distinct from those measured by present subtests, the norms cannot be derived by equating them to the current ASVAB. A new national reference sample will have to be tested. This will make it convenient to define a new reference ASVAB form to replace form 8a.

RECOMMENDATIONS

- There should be only one seeded item per subtest in ACAP.
- The present CAT time limits should be reduced. Time per item equal to that in PP ASVAB is probably adequate.
- The development and validation of new subtests should be considered an integral and important part of the CAT-ASVAB project.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Within a few years the Department of Defense (DOD) may begin administering the Armed Services Vocational Aptitude Battery (ASVAB) using computerized adaptive testing (CAT). Current work on CAT-ASVAB is a part of the Accelerated CAT-ASVAB Project (ACAP). ACAP will provide information about the feasibility of eventual full-scale development (FSD) for nationwide implementation of CAT-ASVAB.

It is possible but unsafe to think of ACAP as a research project which will teach us what will or will not work in FSD. If CAT has to be changed appreciably after ACAP, FSD will have to begin without direct evidence that the new version works properly under operational conditions. Therefore ACAP should be thought of as a pilot project for FSD. Anything of importance which cannot or ought not to be done in FSD should not be done during ACAP.

The purpose of this Research Memorandum is to present the implications of two nontechnical considerations for some technical decisions in ACAP. One is the role of ACAP as the pilot for FSD. The other is how CAT can provide economic benefit for DOD. The outline of the argument is as follows: (i) The ultimate goal is to achieve better selection and classification during FSD. (ii) Substantial improvements can be achieved only by adding new tests to the battery. (iii) Thus the primary role of CAT is to provide satisfactory measurement in less time, so that the time saved can be used for the new tests. (iv) Therefore, in anticipation of FSD, technical decisions in ACAP should be aimed at minimizing the duration of CAT without sacrificing any essentials.

This argument has implications for two major technical issues. One is the number of seeded items, and the other is the duration of each subtest. According to current plans, seeded items will be administered in blocks of 20. Examinees are to be divided into five groups, with different subtests seeded for different groups. Because this procedure creates the risk that the CAT-PP equating may change from one group to another, the topic has been reopened. The present time limits are liberal. Under the current seeding plan, they add up to almost the total testing time of PP ASVAB. The importance of saving time implies that the limits should be reduced.

The major nontechnical implication of the analysis of potential benefits is that the development and validation of new tests are very important to the CAT project.

Two additional technical issues are discussed in this Research Memorandum. One is the feasibility (indeed, the ease) of recalibrating items using CAT data. The other is the control of exposure rates of highly informative CAT items, and its effect on reliability.

Three data sets were analyzed to judge the adequacy of current PP time limits: (i) Navy recruits tested in a research study by the Navy Personnel Research and Development Center (NPRDC) in 1979; (ii) recruits tested under standard conditions for norming ASVAB forms 8, 9 and 10; and (iii) applicants who took form 8a during its initial operational test and evaluation (IOT&E).

IMPROVING SELECTION AND CLASSIFICATION

Adaptive testing is potentially superior to PP testing because it can provide a given level of reliability with fewer items. By using CAT one can save time, obtain more reliable and hence more valid ASVAB scores in the same amount of time, or find a compromise which maximizes total benefit.

Little can be gained by using CAT to increase the reliability of ASVAB. Selection and classification are based on composite rather than subtest scores. As composite reliabilities are already above .9, there is little scope for improving composite validities through more precise measurement. This point is illustrated in table 1 using the four High School Occupational Composites: Mechanical and Crafts (MC), Business and Clerical (BC), Electrical and Electronics (EE), and Health, Social & Technology (HST).

The first two rows in table 1 contain the reliabilities and typical validities in PP ASVAB. (Reliabilities are taken from table 30 in [1]. Validities are average values reported in table 47 in [1].) The next two rows show what they will become if each PP subtest is made longer by a half. It is clear that validities will not increase by much even if CAT precision is equivalent to making PP longer by 50 percent.

TABLE 1

GAINS IN RELIABILITY AND VALIDITY WITH
50 PERCENT INCREASE IN TEST LENGTH

	Composite			
	MC	BC	EE	HST
Reliability: PP	.93	.93	.94	.95
Validity: PP	.48	.45	.48	.49
Reliability: PP+50%	.953	.953	.960	.967
Validity: PP+50%	.486	.456	.485	.494

If CAT is designed to save time while equaling rather than exceeding PP in measurement precision at all ability levels, the time saved can be used to administer new tests which measure skills not measured by the ASVAB. This approach may provide much larger gains in validity, which would lead to better selection and classification of recruits and thus to long-term benefits for DOD.

SEEDED ITEMS

CAT-ASVAB is based on a mathematical model in which all the items for a subtest are assumed to measure a single ability θ. Each item has an item response curve (IRC) which describes how the probability of correctly answering the item increases with ability. The model used in CAT-ASVAB contains three parameters for each item: discrimination (a), difficulty (b), and guessing (c). Items administered during CAT are selected from a large pool using the best available estimate of the examinee's ability. The pool must be calibrated, i.e., parameters of the items must be estimated, before adaptive testing can begin. The item pools to be used in ACAP have already been constructed and calibrated using PP administration [2]. An equating study will be carried out to relate CAT scores to those on ASVAB form 8a and thus to the current ASVAB score scale. (Development of the ASVAB score scale is described in [3].)

The PP forms of ASVAB now in use are changed every four years. Creation of new PP ASVAB forms involves (among other steps) an equating study to relate scores on the new forms to those on the reference form 8a, and an initial operational test and evaluation. The purpose of constructing new forms is to reduce test compromise which can occur if examinees come to know some of the items in the battery.

CAT-ASVAB, too, will require new forms for the same reason. New forms consist of new or modified item pools. However, these item pools can be constructed without special data collection for equating and for evaluation of the equating. Ideally, once CAT-ASVAB is implemented nationwide, new items will be pretested and calibrated by administering them to applicants along with operational CAT items. Introduction of such nonoperational items is called "seeding," and the process of estimating their parameters is called "on-line calibration." Once seeded items have been calibrated, they can be used to replace those in the current operational pool. Considerable savings are expected from not having to collect data for equating every four years.

According to current plans [4,5], seeded items will be administered in blocks of 20 per subtest. The large number of items makes it necessary to divide examinees into five groups. Different groups receive seeded items in different subtests. This procedure creates the risk that the equating of CAT scores to PP scores may change from one group to another. Various alternatives have been considered, but as long as seeded items are to be administered 20 at a time, none of them appears satisfactory [5]. One can return to the earlier plan of administering five seeded items in each subtest to every examinee. This strategy will eliminate equating problems, but will require more testing time. (The current plan with five groups is equivalent to four seeded items per subtest per person.)

A new solution emerges when we consider what would happen in FSD. About a million applicants take the ASVAB every year. (This number is close enough for rough estimates.) Development of a new item pool involves pretesting to select satisfactory items, and then calibration to estimate item parameters. Experience with the ACAP pool [2] indicates that about half the items turn out to be satisfactory. Thus, with 300 examinees per item during the pretest and 3,400 per item for calibration, it takes about 4,000 responses to prepare an item for CAT use. If only one seeded item per subtest is administered to each applicant, it is possible to obtain 250 new calibrated items every year for each subtest.

The ACAP item pool contains 100 items in each of two equivalent forms. Following PP ASVAB, two new forms will be needed every four years. Thus, even with only one seeded item per subtest, on-line calibration in FSD will provide five times the data needed to maintain and replace CAT item pools. Therefore, administering more than one item would be a waste of examinee time which can be put to better use, i.e., used to administer new predictors.

In addition to saving time, use of a single seeded item minimizes the effect of seeding on operational scores. CAT begins with an item of medium difficulty and, for an examinee of low ability, proceeds to administer easier items. Seeded items, in contrast, are administered to all examinees. If seeded items are difficult, they may shake an applicant's self-confidence and hence reduce the operational CAT score. This effect is particularly troublesome if a minority group already has a relatively low mean score on a subtest (e.g. women on information subtests). Such potential disturbance is minimized by keeping the number of seeded items as small as possible.

Factor Analysis

One reason for seeding items in blocks of 20 is to allow for factor analyses of seeded items [4] so that dimensionality of the new items can be examined. However, 20 items are not enough. Factor analyses of the ACAP calibration booklets by NPRDC showed that, even with 33 to 86 items per booklet, the number of factors was not constant from one booklet to another [6]. Even when the number was the same, it was personal judgment and not quantitative analysis which determined that different booklets measured the same factors.

Two operational decisions in ACAP are based on the presence of more than one factor in a subtest. Auto and Shop Information has been split into two CAT subtests, and content balancing will be used to ensure that General Science will contain the correct proportions of Physical Science, Life Science and Chemistry items for each examinee [6]. In both cases the division of items into subtests or categories is based on judgments about their content, not on factor loadings.

Given the importance of subjective judgment and the instability of factor solutions, not much will be gained by factoring 20 items at a time. If and when factor analysis is really needed, data can be collected at recruit training centers.

Scale Drift

Another reason for using large blocks of seeded items lies in the methodology of on-line calibration. According to simulations by Stocking [7], random sampling errors in item parameter estimates lead to systematic errors in parameter estimates of seeded items, causing a drift in the score scale. This problem can be avoided if one can ignore the operational CAT items during on-line calibration. If seeded items are administered in large enough blocks they can be calibrated on their own, as is done with PP booklets, without using responses on the operational CAT items.

Actually the problem is even more serious. Stocking assumed that item parameters did not change in going from the original calibration using PP booklets to computerized adaptive administration. Analysis of the experimental CAT-ASVAB data has shown that this assumption is incorrect [8]. Therefore, before being used for on-line calibration, items in the CAT pool need to be recalibrated using CAT data. Values obtained from recalibration should then be used for estimating parameters of seeded items.

With about a million examinees per year in FSD, even if an item is administered to only 1 percent of applicants, enough data for adequate recalibration will be obtained in three months. It is possible that, even with the same sample size, recalibration will be superior to the original PP calibration in fitting IRC in the ability range where the item tends to be used during CAT. Once the operational items have been recalibrated, sampling errors in the original calibration have no effect on on-line calibration. Whether this procedure solves the problem pointed out by Stocking can be determined from simulations.

Other Considerations

Seeded items can be used to study the effect of the mode of administration (PP vs. CAT) on ability distributions and item parameters, but they are not necessary. Operational CAT items can provide sufficient information. The necessary methodologies are available in [9, 10]. Changes with time in the applicant population, if any, can be studied using operational PP scores during the score equating and IOT&E phases of ACAP.

The one shortcoming of seeding one item per subtest during ACAP results from the fact that ACAP samples will be much smaller than those in FSD. With 5,000 applicants to be administered CAT during score equating and 5,000 during IOT&E, satisfactory on-line calibration can be carried out for only about five items per subtest. This will not provide a convincing demonstration of on-line calibration using real data (although it will show that the necessary computer programs do work). However, if the seeding strategy during ACAP is not to be used during FSD, ACAP will have demonstrated nothing at all about on-line calibration under the FSD strategy.

The discussion above assumes that item recalibration using CAT data is feasible. No algorithm has been published for estimating item parameters from CAT data. In fact, one algorithm which is frequently used with PP data is known not to work with CAT data [11]. However, the approximation developed by the author has worked quite well with the experimental CAT-ASVAB data set [10]. The method has now been extended, and evaluated using simulated data. Results are presented in appendix A. They show that item recalibration is indeed feasible. When a simple approximation works well, maximum likelihood will work even better.

One major aspect of FSD which cannot be tried out during ACAP is the ultimate goal of on-line calibration: developing new CAT forms and making sure that scores on new forms have the same percentile ranks as scores on earlier forms. Therefore the long-term reliability of on-line calibration is not a technical issue in ACAP. However, it is not too early to think about the major aspects of FSD. Therefore appendix B discusses the nature and risks of on-line calibration.

TIME LIMITS

The present time limits for CAT subtests, including seeded items, add up to about the same total testing time as for PP ASVAB (enclosure 3h in [5]). The average amount of time taken by examinees is expected to be appreciably smaller. However, it turns out that this is a problem rather than a benefit. About 85 percent of applicants are tested at small sites which have no rooms or facilities for those who complete the test before the slowest applicant. (In any case it is quite possible that, once CAT scores are operational and affect applicants' careers, the average time will increase. Unlike PP tests, adaptive tests do not allow the examinee to go back and change a previous answer. Hence there is no benefit for the examinee in completing the test in less than the allotted time.) Therefore any time savings must be achieved by having time limits shorter than those of PP forms. The time saved can be used to administer new predictors.

The time limits for power subtests in the PP ASVAB are based on a study conducted by NPRDC on Navy recruits in 1979 [12]. There were three time limits for every subtest. For each time limit the study reported the percentage of recruits whom completed the subtest, and the mean and sigma of number of items reached. There are two differences between the testing conditions in [12] and those encountered during operational use of the ASVAB. One is that applicants are more motivated than recruits to complete the test and score high. The other is that examinees are given different instructions in a research study and in operational administration.

The author has performed two analyses to examine the rate of completion of ASVAB subtests under operational instructions and time limits, using data collected after the NPRDC study. One analysis was based on responses of 2,584 applicants who took form 8a in the 1980 IOT&E. (An examinee was rejected if all items were unanswered on any subtest.) Any unanswered items after the last answered item on a subtest were considered to be "unreached." Distributions of the numbers of unreached items were computed separately for each mental category, except that categories IVb and IVc were combined. (Definitions of mental categories in terms of AFQT percentile scores are given in [3].) The second analysis was similar, but based on the responses of 2,562 recruits who participated in the norming study for forms 8, 9, and 10.

Table 2 contains results of the author's analyses. The percentages of completed subtests and mean numbers of unreached items yield several interesting conclusions. (i) As expected, in any given mental category, applicants complete more tests than recruits do. (ii) From Category I to Category IVa, low-ability examinees leave more items

TABLE 2

COMPLETION RATES AND NUMBERS OF UNREACHED ITEMS ON ASVAB FORM 8A UNDER STANDARD ADMINISTRATION

Sub-	% Completed in category							Mean unreached in category							S.D. unreached in category						
test	I	II	IIIa	IIIb	IVa	IVbc	V	I	II	IIIa	IIIb	IVa	IVbc	V	I	II	IIIa	IIIb	IVa	IVbc	V
Results for applicants: frequencies of mental categories are															56	542	470	479	335	547	155
G S	100.0	99.1	96.4	95.6	90.4	91.6	91.0	0.0	0.0	0.1	0.1	0.3	0.4	0.5	0.0	0.4	0.4	0.6	1.1	1.5	1.9
A R	100.0	94.1	92.6	91.0	91.6	92.9	94.2	0.0	0.2	0.3	0.4	0.4	0.4	0.4	0.0	1.0	1.5	1.7	1.8	1.8	2.2
WK	100.0	99.1	98.5	94.2	88.1	79.9	76.8	0.0	0.0	0.1	0.3	0.7	1.4	1.5	0.0	0.3	0.6	1.4	2.2	3.6	3.7
P C	98.2	97.6	96.8	91.2	89.0	84.1	85.8	0.0	0.0	0.1	0.2	0.3	0.4	0.6	0.1	0.3	0.3	0.8	1.0	1.2	1.9
A S	100.0	98.9	96.4	96.0	89.6	86.3	84.5	0.0	0.0	0.1	0.1	0.6	0.7	0.9	0.0	0.3	0.9	0.7	2.2	2.1	2.9
MK	98.2	97.0	94.0	95.6	97.3	95.2	94.2	0.0	0.2	0.2	0.2	0.1	0.2	0.2	0.1	1.2	1.0	0.9	0.8	1.4	1.6
MC	100.0	94.8	91.3	89.6	86.9	83.9	86.5	0.0	0.2	0.3	0.4	0.5	0.7	0.8	0.0	0.9	1.1	1.4	1.5	2.1	2.7
E I	100.0	98.9	98.1	97.5	92.2	89.6	87.1	0.0	0.0	0.1	0.1	0.3	0.4	0.6	0.0	0.3	0.6	0.5	1.2	1.5	1.8
Results for recruits: frequencies of mental categories are															142	635	763	911	78	33	
G S	98.6	98.1	96.7	91.4	79.5	93.9		0.0	0.0	0.1	0.3	0.8	0.2		0.1	0.2	0.6	1.2	2.0	0.7	
A R	98.6	93.1	90.6	90.7	84.6	100.0		0.0	0.3	0.5	0.5	1.0	0.0		0.3	1.2	2.0	2.0	2.9	0.0	
WK	100.0	99.1	95.4	89.5	79.5	84.8		0.0	0.1	0.2	0.7	1.7	1.2		0.0	0.7	1.3	2.5	4.1	4.0	
P C	100.0	98.4	93.1	89.7	82.1	93.9		0.0	0.0	0.2	0.3	0.6	0.1		0.0	0.3	0.8	1.2	1.7	0.5	
A S	100.0	98.7	95.5	93.0	87.2	90.9		0.0	0.1	0.1	0.4	0.5	0.3		0.0	0.6	0.8	1.9	1.6	1.4	
MK	99.3	95.0	94.0	92.5	89.7	90.9		0.0	0.2	0.3	0.5	0.7	0.8		0.2	1.5	1.5	2.3	2.2	3.6	
MC	100.0	96.5	91.2	88.4	83.3	87.9		0.0	0.1	0.4	0.5	0.7	0.8		0.0	0.8	1.5	1.9	1.9	3.0	
E I	97.9	97.6	95.9	95.0	87.2	97.0		0.0	0.1	0.2	0.2	0.7	0.1		0.2	0.8	1.0	1.0	2.1	0.7	

unreached than high-ability examinees. (iii) Among applicants, even in category IVa, the mean number of unreached items is less than 0.8.

For convenience in making comparisons, table 3 repeats some of the completion rates in table 2 and adds those from the NPRDC study. It presents the three subtests for which one of the NPRDC time limits became the operational ASVAB limit. Mental categories IVa, IVb, and IVc have been merged as in [12]. It is clear from table 3 that recruits were slower in NPRDC's research than under standardized conditions of the 8-9-10 norming study.

It is possible that applicants complete PP ASVAB by not answering some items in the middle, which is not permitted in CAT. Therefore an analysis of omitted items was carried out. An "omit" was defined as any unanswered item preceding the last answered item. Table 4 contains means and standard deviations of numbers of omits. It shows that omits are even less frequent than unreached items, and hence not a matter of concern.

There are several arguments for reducing the present CAT time limits.

- The present CAT limits are based on the performance of Army recruits who took the experimental version of CAT-ASVAB (enclosure 3.2b in [6]). They had nothing to gain by scoring high or by working quickly. The tests had no time limits. Since recruits work faster when the test is timed, and since applicants work faster than recruits, shorter time limits should be satisfactory.

- When a test is not completed, much of the time tends to be spent on the last few items. This tendency is illustrated by the performance of Mental Category IIIa recruits in [12] on General Science. On the average they completed 23.8 items in the first nine minutes, and only 0.7 more in the next two. It is likely that the items at the end of the PP test are too hard for those who fail to complete it, and hence allowing more time might have increased their score only through guessing correctly.

- Items near the end of a CAT subtest are tailored to the examinee's ability and hence, for low ability examinees, are likely to be much easier than items near the end of a PP form. This will reduce the amount of time needed.

- Items are easier to read on the Hewlett Packard computer to be used in ACAP than on the Apple computer used for the experimental version, which should reduce the amount of time per item. This conjecture is

TABLE 3

PERCENTAGES OF TESTS COMPLETED BY RECRUITS IN NPRDC RESEARCH, RECRUITS IN 8-9-10 NORMING, AND APPLICANTS IN IOT&E

Mental	General Science			Auto and Shop Info			Electronics Info		
Category	NPRDC	Norm	IOT&E	NPRDC	Norm	IOT&E	NPRDC	Norm	IOT&E
I	97.6	98.6	100.	98.8	100.	100.	98.8	97.9	100.
II	98.4	98.1	99.1	95.6	98.7	98.9	97.1	97.6	98.9
IIIa	92.2	96.7	96.4	86.5	95.5	96.4	92.5	95.9	98.1
IIIb	84.2	91.4	95.6	77.3	93.0	96.0	86.4	95.0	97.5
IV	82.9	83.8	91.1	68.6	88.3	87.6	88.6	90.1	90.6

TABLE 4
ANALYSIS OF NUMBERS OF OMITS ON ASVAB FORM 8A UNDER STANDARD ADMINISTRATION

Sub-	Mean omits in category							S.D. of omits in category						
test	I	II	IIIa	IIIb	IVa	IVbc	V	I	II	IIIa	IIIb	IVa	IVbc	V
					Results for applicants:									
	Frequencies of mental categories are							56	542	470	479	335	547	155
G S	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.0	0.4	0.2	0.5	0.8	0.6	0.9
A R	0.0	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.8	0.7	0.7	0.6	0.6	0.4
W K	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.0	0.4	0.2	1.2	1.2	0.9	0.8
P C	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.5	0.8
A S	0.0	0.1	0.0	0.1	0.1	0.1	0.2	0.3	0.6	0.3	0.5	0.9	0.9	1.0
M K	0.1	0.2	0.3	0.3	0.3	0.2	0.2	0.5	0.9	1.4	1.5	1.6	1.0	1.4
M C	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.0	0.4	0.3	0.3	0.4	0.9	1.1
E I	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.3	0.3	0.2	0.3	0.3	0.8
					Results for recruits:									
	Frequencies of mental categories are							142	635	763	911	78	33	
G S	0.0	0.0	0.0	0.1	0.2	0.2		0.0	0.2	0.3	0.8	1.3	0.6	
A R	0.1	0.1	0.2	0.3	0.3	0.2		0.6	0.7	0.8	1.5	0.9	0.6	
W K	0.0	0.0	0.1	0.2	0.3	0.2		0.1	0.3	0.7	0.9	0.8	0.4	
P C	0.0	0.0	0.0	0.1	0.1	0.1		0.0	0.1	0.2	0.3	0.2	0.3	
A S	0.0	0.0	0.1	0.1	0.0	0.1		0.4	0.3	0.6	0.8	0.2	0.4	
M K	0.1	0.2	0.3	0.4	0.2	0.2		1.1	1.0	1.5	1.7	0.7	0.6	
M C	0.0	0.1	0.1	0.1	0.0	0.1		0.2	0.4	0.4	0.6	0.2	0.2	
E I	0.0	0.0	0.0	0.1	0.1	0.0		0.2	0.4	0.4	0.5	0.5	0.2	

supported by the data on which the CAT limits are based. The 10-item Paragraph Comprehension test in the experimental CAT was completed by 95 percent of Army recruits in 16.8 minutes (enclosure 3.2b in [6]). The PP ASVAB limit for PC amounts to 8.7 minutes for 10 items. The ratio of these durations is 1.84. When similar ratios are calculated for all other subtests, none is higher than 1.38. The larger ratio for PC is attributable to the fact that it requires much more reading than other subtests.

The current per-item time limits for CAT subtests were chosen by the CAT-ASVAB Psychometric Committee [6]. If one item is seeded in each subtest, and time limits for individual subtests are not rounded to integers, the limits for power subtests add up to 108.4 minutes. If one allows the same amount of time per item as in PP ASVAB, the total testing time will be 85.2 minutes. Thus, 23 minutes can be saved by replacing the current time limits by those proportional to PP ASVAB limits which, according to table 2, are quite reasonable.

The choice of time limits depends on policy considerations as well as technical ones. Psychometricians can analyze data on completion rates, numbers of unreached items, etc. However, converting this information into time limits for CAT involves value judgments. The decision depends on what will be done with the time saved by replacing PP ASVAB with CAT.

IMPORTANCE OF NEW PREDICTORS

As illustrated in table 1, the accuracy of selection and classification cannot be improved substantially except by adding new predictors to the battery. Therefore the value of CAT to DOD depends heavily on the incremental validities of the new subtests, i.e., on the extent to which they increase the validities of the AFQT and of other composites. Values of these increments are fundamental to any economic analysis of the benefits of CAT. Therefore, development and validation of new tests are an integral and important part of the CAT-ASVAB project.

Even after research is complete and the incremental validities of some new tests have been judged satisfactory, adding them to the ASVAB battery is likely to be a long process. The reason is that each service has its own composites, and within a service there are different composites for different occupational specialties. Any composite which can be improved by addition of a new test will have to be redefined. If the new definitions have to be justified on the basis of their predictive validities, the data collected upto that time may not suffice; new data may have to be collected. To avoid unnecessary delays, the course of this entire process should be anticipated, and plans made accordingly. Different predictors are being developed independently by different services. Close cooperation among the services will be necessary for a smooth transition to a new ASVAB in a reasonable amount of time.

As the analyses of test completion rates (tables 2 and 3) have shown, results obtained under research conditions may not remain valid under operational conditions. In addition, one must consider sensitivity to coaching and practice. A new predictor will increase validity to the extent that it measures a skill not measured by current ASVAB subtests. If speed of response is a significant component of this skill, it is likely that the test score can be improved through practice and by using the proper strategy. While research subjects will not seek coaching to improve their scores in this manner, many applicants will. Therefore the choice of additional subtests may present a dilemma; the tests with the highest incremental validities may also be the ones most susceptible to coaching.

Introduction of new tests, especially new kinds of tests, will require testing of a new nationwide reference sample. Because they will measure quite different traits from the present ASVAB subtests, norms for new predictors cannot be derived by equating them to current subtests.

EXPOSURE CONTROL AND RELIABILITY

If CAT items are selected for maximum information, without any constraints, highly discriminating items of medium difficulty are administered very frequently. This increases the risk of test compromise through examinees telling others about the items they remember. This risk will be reduced in ACAP by controlling the exposure rates of items [4, 13]. The exposure control parameters are calculated using simulations. Their values are such that a given item will not be administered to more than one-sixth of examinees, which is the exposure rate of PP items when six distinct forms are in use at one time.

It should be remembered that item exposure rate is controlled as an average over the entire national population. Applicants coming to a specific recruiter will have a much smaller range of ability, in which case items suited for this range may be administered to this group more than one-sixth of the time. Nothing can be done about this problem, except to develop measures for detecting test compromise which may be limited to a single recruiter or a small area.

The chance of test compromise depends not only on exposure rate per item but also on test length [14]. Suppose only one item is administered to each examinee.

With exposure rate controlled at one-sixth, if examinees know the six commonly used items, they are almost certain of being administered one of them. (The certainty is not complete because random numbers are used in item selection under exposure control.) Thus, with any given level of exposure control, a shorter test is more susceptible to compromise. (It is possible that this risk is not a major concern because very few recruiters try to cheat [15].)

The two major influences on CAT reliability are test length and exposure control. Exposure control reduces the use of the most informative items and hence lowers reliability. Simulations have shown that, with item pools to be used in ACAP, CAT scores on General Science are less precise than PP scores at medium abilities (enclosure 7 in [5]). It is almost certain that these simulations overestimate CAT precision. Analysis of the experimental CAT-ASVAB has shown that item parameters change from PP to CAT administration [8]. Therefore the use of estimates obtained from PP data constitutes mis-specification of IRCs and hence reduces reliability.

For General Science, which is not a part of the Armed Forces Qualification Test (AFQT), the current (tentative) solution is to raise the exposure rate [5]. This approach may not be acceptable if the same problem occurs with a subtest in the AFQT, and an increase in test length may become necessary.

RECOMMENDATIONS

- One seeded item per subtest should be used in FSD, and hence during ACAP.

- The present time limits for CAT subtests should be reduced. Time per item in PP ASVAB seems to be adequate, and hence it can be used to set CAT time limits.

- Development and validation of new tests should be considered an integral and important part of the CAT-ASVAB project. The services should begin discussions for avoiding unnecessary delays in the long process of adding new tests to the operational ASVAB and of redefining composites.

- Before test lengths and exposure rates are given final approval, reliabilities of CAT subtests should be examined using simulations which include the effect of the mode of administration.

REFERENCES

[1] U.S. Military Entrance Processing Command, Technical Supplement to the Counselors' Manual for the Armed Services Vocational Aptitude Battery Form 14. North Chicago: U.S. Military Entrance Processing Command, 1985

[2] Air Force Human Resources Laboratory, TR-85-19, "Armed Services Vocational Aptitude Battery: Development of an Adaptive Item Pool," by J. Stephen Prestwood, C. David Vale, Randy H. Massey, and John R. Welsh, Sep 1985

[3] CNA, Report 116, "Constructing the 1980 ASVAB Score Scale," by Milton H. Maier and William H. Sims, Jul 1986

[4] CNA, Memorandum 86-0442, "Minutes of November 1985 Meeting of the CAT-ASVAB Psychometric Committee," by William H. Sims, 13 Mar 1986

[5] Naval Postgraduate School, Memorandum, "Minutes of September 1986 Meeting of the CAT-ASVAB Psychometric Committee," by Bruce Bloxom, 17 Oct 1986

[6] Naval Postgraduate School, Memorandum, "Minutes of March 1986 Meeting of the CAT-ASVAB Psychometric Committee," by Bruce Bloxom, 23 May 1986

[7] Stocking, Martha. "A Progress Report on a LOGIST-based On-line Calibration Method," presentation at ONR Contractors' Meeting on Model-Based Psychological Measurement, 30 Apr 1986

[8] CNA, Research Memorandum 86-24, "Effect of the Medium of Administration on ASVAB Item Response Curves," by D. R. Divgi and Peter H. Stoloff, Apr 1986

[9] Mislevy, Robert J. "Estimating Latent Distributions," Psychometrika(Dec 1984): 359-381

[10] CNA, Research Memorandum 86-189, "Sensitivity of CAT-ASVAB Scores to Changes in Item Response Curves with the Medium of Administration," by D. R. Divgi, Aug 1986

[11] Lord, Frederic M., Educational Testing Service, letter to J. B. Sympson, Navy Personnel Research & Development Center, regarding parameter estimation using adaptive test data, 15 Oct 1984

[12] Navy Personnel Research and Development Center, Memorandum, "Recommended Time Limits for ASVAB 8a Power Tests," by Martin F. Wiskoff, 14 Dec 1979

[13] Sympson, James Bradford and Hetter, Rebecca Danon, "Controlling Item-Exposure Rates in Computerized Adaptive Testing," presented at the Annual Conference of the Military Testing Association, 21 Oct 1985

[14] Earles, James. Comments during the September 1986 meeting of the CAT-ASVAB Psychometric Committee

[15] CNA, Report 110, "Extent of Cheating on ASVAB, FY 1976 Through FY 1984," by William H. Sims, Ann R. Truss, and Marjorie D. Curia, Nov 1985

APPENDIX A
FEASIBILITY OF ITEM RECALIBRATION WITH CAT DATA

No algorithm has been published for estimating item parameters from CAT data. In fact, one algorithm which is frequently used with PP data is known not to work with CAT data [A-1]. However, the approximation developed by the author has worked quite well with the experimental CAT-ASVAB data set [A-2].

The major question is whether the guessing parameter c can be estimated adequately, and if not, whether the errors in IRC are unimportant in the ability range where the item tends to be administered during CAT. Therefore a simple procedure, based on closed expressions for discrimination a and difficulty b [A-2], was examined using simulated data. The 100-item pool was the one used by Davis [A-3] to simulate the Word Knowledge subtest. It was picked by Davis from the ACAP item pool [A-4]. A sample of 15,000 examinees was generated from a standard normal distribution of ability. Exposure rates of the items were controlled using the procedure described by Davis [A-3].

Two sets of parameter estimates were obtained for each item administered to at least 1,500 examinees. One was calculated by setting c equal to zero to see whether, in spite of such oversimplification, estimated IRCs remain satisfactory at the ability levels of interest. The second set of estimates was obtained by partial maximum likelihood. For each value of c, a and b were calculated as in [A-2], which limited the maximization to one dimension instead of three. The estimates were used to compute the likelihood of the observed proportions of correct answers in 61 ability groups in the interval from -3 to 3. The estimate of c was obtained by maximizing the likelihood function.

Table A-1 shows results for items administered to at least 1,500 examinees. The item code in the first column is the position of the item in the ACAP pool during calibration [A-4]. The second column shows the number of examinees to whom the item was administered. Columns 3 and 4 contain the mean and standard deviation of the estimated abilities of these examinees. Columns 5 to 7 present the parameters to generate the simulated tests. The next two columns show the estimates of discrimination (a) and difficulty (b) when the guessing parameter was set equal to zero, followed by the average absolute deviation (AAD) of the resulting IRC from the true IRC. (The average was computed over the examinees who received the item, using their estimated abilities.) The last four columns contain the values obtained when c was estimated by partial maximum likelihood, and the AAD of the corresponding IRC.

TABLE A-1

RESULTS OF RECALIBRATION USING SIMULATED CAT DATA

		Estimated θ		Parameters			Estimates with c=0			Estimates with fitted c			
Item	N	Mean	S.D.	a	b	c	a	b	AAD	a	b	c	AAD
1	3627	0.03	0.37	1.54	-0.11	.29	0.95	-0.54	.021	1.10	-0.29	.18	.017
7	3907	0.64	0.51	1.69	0.80	.07	1.39	0.73	.016	1.81	0.83	.10	.007
9	2918	0.66	0.36	1.38	0.69	.13	1.09	0.56	.013	1.17	0.62	.05	.013
28	4079	-0.52	0.35	1.52	-0.77	.21	1.20	-1.00	.010	1.38	-0.80	.19	.009
31	1621	-1.48	0.39	1.83	-1.75	.13	1.49	-1.84	.022	1.65	-1.74	.10	.021
34	3824	-0.89	0.42	1.56	-1.20	.12	1.36	-1.33	.009	1.53	-1.19	.13	.004
35	4730	0.12	0.56	1.55	0.04	.15	1.15	-0.14	.022	1.41	0.03	.14	.009
41	2273	1.41	0.44	1.89	1.57	.14	1.40	1.45	.019	2.04	1.61	.17	.012
42	2962	-0.01	0.89	2.07	-0.25	.12	1.49	-0.40	.032	1.80	-0.29	.09	.013
45	3875	-0.48	0.54	2.21	-0.64	.22	1.48	-0.90	.036	2.20	-0.64	.23	.006
48	2203	-0.93	0.29	1.62	-1.25	.27	1.56	-1.43	.009	1.98	-1.15	.33	.011
54	4024	0.53	0.48	1.69	0.56	.20	1.15	0.33	.026	1.64	0.57	.20	.005
62	1614	-1.42	0.31	1.72	-1.64	.25	1.55	-1.84	.010	1.83	-1.66	.21	.009
73	4448	-0.11	0.59	1.75	-0.24	.19	1.20	-0.48	.030	1.55	-0.26	.18	.013
77	4938	0.05	0.44	1.52	-0.07	.20	1.04	-0.33	.023	1.29	-0.08	.19	.016
79	3887	0.43	0.40	1.61	0.43	.25	0.91	0.08	.032	1.11	0.33	.17	.026
81	4016	-0.04	0.66	2.04	-0.20	.27	1.05	-0.60	.053	1.50	-0.23	.26	.030
87	1688	1.50	0.26	1.93	1.46	.30	1.36	1.24	.034	2.46	1.63	.40	.036
88	5067	0.50	0.48	1.38	0.65	.04	1.17	0.62	.014	1.28	0.68	.05	.011
90	3718	0.74	0.47	1.79	0.81	.11	1.44	0.75	.018	1.99	0.89	.15	.022
91	3583	-0.43	0.33	1.45	-0.67	.25	1.08	-1.02	.019	1.09	-1.00	.02	.018
93	3894	0.28	0.64	1.67	0.33	.11	1.19	0.21	.030	1.75	0.41	.16	.016
109	4154	-0.04	0.39	1.27	-0.10	.10	1.09	-0.24	.009	1.12	-0.21	.03	.008
111	3045	1.01	0.40	1.86	1.03	.18	1.27	0.85	.024	1.46	0.97	.12	.018
116	3142	0.52	0.35	1.47	0.51	.21	1.10	0.28	.011	1.42	0.52	.21	.007
118	3072	-0.81	0.51	1.81	-1.12	.10	1.67	-1.20	.008	1.72	-1.18	.02	.006
121	1547	1.64	0.42	2.08	1.78	.18	1.22	1.62	.035	1.40	1.73	.10	.029
123	4396	-0.27	0.36	1.50	-0.40	.22	1.11	-0.66	.013	1.17	-0.59	.06	.012
127	4204	-0.64	0.39	1.66	-0.87	.22	1.31	-1.11	.015	1.68	-0.85	.25	.004
128	5123	-0.31	0.43	1.47	-0.46	.10	1.28	-0.57	.008	1.41	-0.47	.09	.004
129	4429	-0.36	0.56	1.78	-0.51	.12	1.39	-0.68	.024	1.61	-0.56	.11	.013
136	3327	-0.09	0.78	2.20	-0.28	.25	1.24	-0.62	.049	2.37	-0.23	.28	.010
137	4423	-0.51	0.51	1.67	-0.67	.12	1.31	-0.81	.020	1.42	-0.74	.06	.015
152	2050	-0.88	0.25	1.91	-1.15	.45	1.63	-1.47	.008	2.16	-1.14	.40	.012
162	3960	0.27	0.35	1.49	0.26	.24	1.09	0.00	.011	1.37	0.23	.19	.008
165	3820	-0.04	0.73	1.99	-0.21	.22	1.18	-0.51	.044	1.52	-0.28	.17	.024
170	3241	-0.87	0.44	1.82	-1.13	.25	1.32	-1.43	.022	1.54	-1.24	.17	.012
172	2756	-0.69	0.35	1.35	-1.13	.12	1.39	-1.17	.014	1.48	-1.07	.11	.015
174	2602	1.10	0.42	1.94	1.08	.22	1.27	0.88	.024	2.49	1.24	.34	.032
176	4187	-0.15	0.34	1.89	-0.30	.43	1.07	-0.86	.022	1.54	-0.35	.40	.012
178	3386	0.64	0.54	2.32	0.71	.25	1.17	0.43	.053	1.92	0.73	.25	.021
179	3969	0.42	0.55	1.84	0.50	.22	1.07	0.24	.039	1.68	0.53	.23	.014
183	3906	0.80	0.41	1.60	0.89	.07	1.32	0.83	.013	1.62	0.94	.11	.010
187	3416	0.93	0.38	1.70	0.97	.13	1.40	0.83	.017	1.51	0.88	.06	.013
188	2572	1.19	0.37	1.79	1.23	.19	1.27	1.06	.017	1.80	1.29	.22	.014
194	4935	-0.09	0.46	1.58	-0.22	.19	1.05	-0.50	.028	1.52	-0.14	.28	.018
217	4386	0.05	0.53	1.82	-0.10	.33	1.03	-0.55	.034	1.70	-0.05	.35	.013
224	2204	-0.93	0.39	1.33	-1.31	.06	1.34	-1.35	.005	1.34	-1.35	.00	.005
226	1790	1.37	0.30	1.52	1.42	.16	1.28	1.30	.025	1.49	1.42	.12	.024
229	2041	-1.30	0.41	1.89	-1.54	.17	1.61	-1.73	.023	1.68	-1.68	.05	.021
230	3423	-0.97	0.44	1.58	-1.25	.09	1.53	-1.32	.007	1.55	-1.31	.01	.007
235	4146	0.33	0.37	1.42	0.34	.17	1.01	0.18	.023	1.51	0.49	.24	.026
237	4238	-0.54	0.38	2.00	-0.71	.38	1.26	-1.17	.026	1.80	-0.77	.35	.009
240	2167	0.81	0.35	1.33	0.81	.14	1.16	0.64	.010	1.16	0.64	.00	.010
242	2742	0.79	0.53	2.28	0.99	.07	1.72	0.94	.025	2.09	0.99	.07	.009
244	2102	1.37	0.31	2.01	1.39	.31	1.21	1.17	.048	2.30	1.55	.37	.050
245	3342	0.28	0.71	2.15	0.42	.10	1.45	0.34	.039	2.23	0.47	.12	.016
253	3135	-0.90	0.37	1.88	-1.16	.29	1.47	-1.43	.013	1.89	-1.17	.27	.004
254	2593	-1.10	0.41	1.98	-1.30	.27	1.38	-1.60	.023	2.27	-1.22	.35	.010
258	2069	-1.12	0.32	1.69	-1.35	.27	1.45	-1.62	.017	1.47	-1.59	.03	.017

It is clear that the mean and standard deviation of ability change from one item to another. The mean has a strong correlation with the difficulty parameter of the item, which is a consequence of adaptive item selection. When c is estimated the mean of AAD over all items is 0.014, which shows that the estimation is satisfactory. IRCs obtained by setting c equal to zero are worse but acceptable, the mean AAD being 0.023.

The bottom line for on-line calibration is whether ability estimates computed after recalibration are sufficiently accurate. Therefore three ability estimates were calculated for each examinee using Owen's approximation [A-5]. These were based, respectively, on the true item parameters, item parameter estimates obtained with c=0, and estimates obtained by partial maximum likelihood. (Since no recalibration was performed for items with sample size below 1,500, the true parameters were used. However, such items accounted for less than 10 percent of the responses in the data set.) The corresponding root-mean-squared difference from true abilities increased by less than 0.001 (i.e., less than 0.01 on the standard score scale) when item parameters were replaced by estimates. Clearly the additional error in ability estimates is negligible.

Results of recalibration are bound to be even better when discrimination and difficulty parameters are estimated by maximum likelihood rather than an approximation. The methodology can be developed and tested well before the end of ACAP. Therefore it can be used to recalibrate operational ACAP items as a prelude to on-line calibration of the seeded items.

REFERENCES

[A-1] Lord, Frederic M., Educational Testing Service, letter to J. B. Sympson, Navy Personnel Research & Development Center, regarding parameter estimation using adaptive test data, 15 Oct 1984

[A-2] CNA, Research Memorandum 86-189, "Sensitivity of CAT-ASVAB Scores to Changes in Item Response Curves with the Medium of Administration," by D. R. Divgi, Aug 1986

[A-3] Office of Naval Research, Memorandum 86-345, "Online Calibration Update," by Charles E. Davis, 11 Sep 1986

[A-4] Air Force Human Resources Laboratory, TR-85-19, "Armed Services Vocational Aptitude Battery: Development of an Adaptive Item Pool," by J. Stephen Prestwood, C. David Vale, Randy H. Massey, and John R. Welsh, Sep 1985

[A-5] Owen, Roger J. "A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing," Journal of the American Statistical Association(Jun 1975): 351-356

APPENDIX B
NATURE AND RISKS OF ON-LINE CALIBRATION

Successful on-line calibration during FSD will make it unnecessary to re-equate CAT to PP scores every time the CAT item pool is replaced or changed, saving a lot of time and effort. However, success is not guaranteed at present. Evaluation of the methodology using simulated data is not yet complete. What works with simulated data may or may not work equally well with real data. The main danger is a slow but steady drift in the CAT score scale which, after some years, makes the operational norm tables invalid.

To put this problem in perspective, let us compare the processes of developing new forms in PP and CAT ASVAB. The item writing stage is the same for both. Tryouts of PP items are conducted by testing recruits. CAT ASVAB can use applicants for this purpose by seeding the new items in the operational test. Small samples, about 300 per item, suffice for item tryout. Following [B-1], about half the items will be good enough to be included in the new forms. Then comes the fundamental difference between PP and CAT.

For PP ASVAB the next step is to collect equating data by administering the new forms along with the reference form 8a. Score distributions are used to equate each new form to 8a. As with CAT items, characteristics of PP items can change with time due to changes in the applicant population (e.g. educational level, emphasis in course work, etc.). There are two reasons why such changes do not cause problems. One is that new forms are constructed to be as similar as possible to 8a. Any factors which affect scores on new forms have the same effect on 8a scores. Therefore the equating relationship between the two is robust. The other reason is that, after four years, the new forms are equated directly to 8a and not indirectly through the forms which are about to be replaced. This prevents accumulation of errors from one generation of forms to the next.

In case of CAT only the forms used at the beginning are equated directly to 8a. Afterwards the new items are calibrated on-line, and the fundamental assumption is made that even if the item pool changes, the ability scale of the model remains the same. Therefore there is no need to equate new CAT forms directly to anything.

If the assumption of invariance is correct, on-line calibration is valuable because it will save DOD the expense and effort of data collection and analysis every four years. If the assumption is incorrect, on-line calibration is unsafe. The two safety features of the PP process are absent. Since CAT is not parallel to the reference form 8a, drift in CAT scores may not be cancelled by the same drift in 8a scores. As new forms are equated to 8a only indirectly, through the current CAT forms, the drift can accumulate.

All assumptions are wrong to some extent, including those underlying the model on which CAT-ASVAB is based. Two assumptions which are known to be wrong are the three-parameter logistic form of IRCs, and unidimensionality of the item pool of a given subtest. Therefore there is no guarantee that scale drift will not occur. This does not mean, however, that on-line calibration is useless. Stability checks and corrective actions are available.

(i) As in ACAP IOT&E, CAT and form 8a can be administered to random subgroups in some locations. This will provide a direct check on scale drift, but may be considered disruptive. (ii) The same thing can be done using recruits, but then we have just the kind of study which on-line calibration is supposed to eliminate. (iii) As Sims [B-2] has pointed out, one of the two original CAT item pools should be left intact, so that its score scale remains unchanged, and used instead of 8a in (i). This will be unobtrusive and inexpensive. However, it will not detect any drifts which affect CAT item pools differently from PP forms. Therefore it is important that, as long as PP form 8a is considered the reference ASVAB test, CAT items be similar to 8a items in content and format.

If it is found that scores on the current CAT forms are not on the same scale as the reference form, a new equating study will be needed. If a CAT form rather than 8a becomes the reference form, the data collection recommended by Sims [B-2] will serve as the equating study.

If new tests are added to the ASVAB, and these measure traits quite distinct from those measured by the present subtests, norms for the new tests cannot be derived by equating them to the existing subtests. A new national reference sample will have to be tested, at which time it will be a relatively simple matter to redefine the reference form.

REFERENCES

[B-1] Air Force Human Resources Laboratory, TR-85-19, "Armed Services Vocational Aptitude Battery: Development of an Adaptive Item Pool," by J. Stephen Prestwood, C. David Vale, Randy H. Massey, and John R. Welsh, Sep 1985

[B-2] Naval Postgraduate School, Memorandum, "Minutes of March 1986 Meeting of the CAT-ASVAB Psychometric Committee," by Bruce Bloxom, 23 May 1986, enclosure (3.10a).

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) CRM 86-231	5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION Center for Naval Analyses	6b. OFFICE SYMBOL *(If applicable)* CNA	7a. NAME OF MONITORING ORGANIZATION Commandant of the Marine Corps (Code RDS)
6c. ADDRESS *(City, State, and ZIP Code)* 4401 Ford Avenue Alexandria, Virginia 22302-0268		7b. ADDRESS *(City, State, and ZIP Code)* Headquarters, Marine Corps Washington, D.C. 20380
8a. NAME OF FUNDING / ORGANIZATION Office of Naval Research	8b. OFFICE SYMBOL *(If applicable)* ONR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-83-C-0725
8c. ADDRESS *(City, State, and ZIP Code)* 800 North Quincy Street Arlington, Virginia 22217		10. SOURCE OF FUNDING NUMBERS

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
65153M	C0031		

11. TITLE *(Include Security Classification)*

On Some Issues in the Accelerated CAT-ASVAB Project

12. PERSONAL AUTHOR(S)
D. R. Divgi

13a. TYPE OF REPORT Final	13b. TIME COVERED FROM TO	14. DATE OF REPORT *(Year, Month, Day)* November 1986	15. PAGE COUNT 36

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The Accelerated CAT-ASVAB Project (ACAP) may serve as the pilot version for national implementation of computerized adaptive testing (CAT) for the Armed Services Vocational Aptitude Battery (ASVAB). Two major decisions in ACAP involve the introduction of new items into the tests, and setting time limits. This Research Memorandum takes the position that the long-term benefits which CAT may provide are more important than purely technical concerns and makes recommendations based on this position.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT □ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt.Col. G. W. Russell	22b. TELEPHONE *(Include Area Code)* (202) 694-3491	22c OFFICE SYMBOL RDS-40

83 APR edition may be used until exhausted.
All other editions are obsolete.